Trustworthy Artificial Intelligence: A Review

Year: 2022 | Citations: 486 | Authors: Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, A. Durresi

---

Abstract

Artificial intelligence (AI) and algorithmic decision making are having a profound impact on our daily lives. These systems are vastly used in different high-stakes applications like healthcare, business, government, education, and justice, moving us toward a more algorithmic society. However, despite so many advantages of these systems, they sometimes directly or indirectly cause harm to the users and society. Therefore, it has become essential to make these systems safe, reliable, and trustworthy. Several requirements, such as fairness, explainability, accountability, reliability, and acceptance, have been proposed in this direction to make these systems trustworthy. This survey analyzes all of these different requirements through the lens of the literature. It provides an overview of different approaches that can help mitigate AI risks and increase trust and acceptance of the systems by utilizing the users and society. It also discusses existing strategies for validating and verifying these systems and the current standardization efforts for trustworthy AI. Finally, we present a holistic view of the recent advancements in trustworthy AI to help the interested researchers grasp the crucial facets of the topic efficiently and offer possible future research directions.